AWS MACHINE LEARNING SPECIALTY: HOW I GOT CERTIFIED IN TEN DAYS AS A DEVOPS ENGINEER

APN Training Partner | AWS Training and Certification | ML

Alessandro Gaggia | 10 July 2020

Introduction

He did it!

Our Cloud Expert Alessandro Gaggia got his sixth (!) AWS Certification (the 58th AWS Certification for beSharp): the **AWS Certified Machine Learning Specialty!**

Alessandro is considered a backbone of our company: he joined the team as a Front-end developer back in 2012, a few months after beSharp's establishment. Today, he is the dean of our development team. He first approached Machine Learning one year ago while participating in the AWS DeepRacer League held during the AWS RE: Mars in Las Vegas. From that moment onwards, it became a real passion for him. After the rush, which took him towards the certification, Alessandro threw himself into a precious retrospective study on the winning path that led him to obtain this excellent result in a few days. So it's time to leave the floor. Ready to become AWS Certified?

Many congrats to our #CloudExpert @Balubor! 🎉
So #Proud2beCloud! #AWSCertified #MachineLearning #AWS @awscloud @AWSCertifiedBot https://t.co/viMskr5nL1

— beSharp (@beSharpsrl) June 8, 2020

My personal experience

AWS Certified Machine Learning – Specialty is an advanced certification a bit different from the others, because it is the only one which focuses on specific sector knowledge not strictly tied to AWS services. In fact, in order to pass the exam and obtain the certification, it's fundamental being able to **recognize, analyze and optimize different machine learning problems** starting from use cases' descriptions, without them being exclusively linked to peculiar AWS' solutions.

This is my point of view and it diverges a little from what I have read in many articles, where authors consider SageMaker as the main focus of the exam. I believe, instead, that this exam tested mainly me being familiar with Machine Learning in general.

On the other hand, many testimonials – given by expert data scientists – say that, before taking this certification, many years of on-the-field training are required. Personally, I don't believe this is entirely true: experience always plays a fundamental role however, like many other certifications I've obtained, what you really need is dedication to study, a strong desire to deep dive in all the arguments and logic ability to make it through. Also some understanding of Big Data concepts can help.

Disclaimer: information gathered here repòresent my personal experience on preparing the exam and it is by no means to be intended as a substitute, more over exhaustive, of studying material proposed by Amazon Web Services. I believe though, that they are an excellent reference material to follow in order to optimize your learning path, containing references to many different sites about Machine Learning and on AWS' FAQ.

Be also extra careful about the different **questionnaires available for testing**, for which you'll get the link, in particular the Certbolt executable, which allows to simulate a complete test of 65 questions with solutions.

With a study full immersion session of about 8 hours a day, certification can be achieved in 1 days... or at least, this was my strategy, which gave me the ability to pass the exam with a score of 800/1000 on a minimum of 750/1000.

Exam's topics

The exam is **170 minutes** long for **65 questions**, personally, I managed to finish in 9 minutes, which gave me plenty of extra time for checking. One positive note is that, unlike other advanced certifications, this one has **short use cases, which means less time to read and more time to think.**

I've put a list of all the main topics i found in both exam readiness test and in the official one:

- Given a ML problem, use a combination of AWS managed services to solve it in the most fast and efficient way possible. In this case we have the following topics:
  - **Kinesis Data Stream, Analytics, Firehose.**
  - **S3, KMS, DynamoDB** (just remember what they are and what they do)
  - **AWS ElasticSearch con Kibana, Splunk, AWS Quicksight** (for Business intelligence – a.k.a **BI**)
  - **AWS Athena, AWS Redshift**
  - **AWS EMR with Spark**
  - **AWS Glue**
  - **AWS Translate, Transcribe, Lex, Polly, DeepLens, Rekognition, Comprehend, etc.**

In some cases we also have a combination of Managed Services and SageMaker, but in these specific cases it's usually trivial to exclude wrong answers.

- Given a ML problem, being able to manage the following scenarios:
  - Data cleaning with appropriate techniques like: **Filtering, Transforming, Scaling, Extracting**
  - Manage **Overfitting/Underfitting**, understanding when we have to work with: **Learning Rate, Batch Size, Oversampling, Regularization, Dropout Rate, Feature Increasing, Denoising, Normalization, Epoch Time, etc.**
  - Evaluate the appropriate score for a specific model, keeping in mind its typology and Business requests; learn how to choose between: **RMSE, R2, F1, Recall, Accuracy, Precision, AUS-ROC, etc.**
  - Manage **Hyperparameters** tuning both on Sagemaker and in general; you need to see: **Learning Rate, Epoch Time, L1 e L2 Regularization, Adam, Random and Stocastic Optimization, alpha, beta** and **gamma parameters** in different contexts.
  - **Manage deploy on Sagemaker.**

To resume all the projects in details, some "must" are:

- Creation of a pipeline using Kinesis Stream, Firehose and Analytics in combination with Amazon Athena or Elastic Search

- Solve a Overfitting problem
- Solve a Tuning problem
- Solve a Scoring problem
- Solve a Confusion matrix problem
- Use a combination of managed AWS ML services to solve a problem in a easy and efficient way
- **Manage different deploy scenarios on Sagemaker**

This list covers the majority of questions you'll likely to see in an exam. Following I will propose a study guide, making it possible to manage all the topics in a tidy and organic way.

Topics to study: a lineup

Reference links about topics here described are mostly from AWS official documentation, in particular for SageMaker and others managed services. Also, regarding topics more related to Machine Learning, I want to propose some sites that I found particularly complete, well explained and exhaustive.

- What is machine learning: how is different from Artificial Intelligence and Deep Learning:
  https://machinelearningmastery.com/machine-learning-for-programmers/
  https://www.geeksforgeeks.org/difference-between-artificial-intelligence-vs-machine-learning-vs-deep-learning/?ref=rp

- ML pipeline in general and on AWS: https://medium.com/slalom-data-analytics/mlops-part-2-machine-learning-pipeline-automation-with-aws-1ca10348239e
- How we proceed on managing a ML problem: https://towardsdatascience.com/how-to-approach-a-machine-learning-problem-3fe843fd1166https://towardsdatascience.com/task-cheatsheet-for-almost-every-machine-learning-project-d0946861c6d0

- Machine Learning on AWS introduction: Managed Services (focus on offered services)
  - AWS GLUE: https://aws.amazon.com/it/glue/faqs/
  - AWS DATAPIPELINE: https://aws.amazon.com/it/datapipeline/faqs/
  - AWS AUGMENTED AI: https://aws.amazon.com/it/augmented-ai/faqs/
  - AWS DEEP LENS: https://aws.amazon.com/it/deeplens/faqs/
  - AWS TRANSLATE: https://aws.amazon.com/it/translate/faqs/
  - AMAZON TRANSCRIBE: https://aws.amazon.com/it/transcribe/faqs/
  - AMAZON TEXTRACT: https://aws.amazon.com/it/textract/faqs/
  - AWS REKOGNITION: https://aws.amazon.com/it/rekognition/faqs/
  - AWS POLLY: https://aws.amazon.com/it/polly/faqs/
  - AWS PERSONALIZE: https://aws.amazon.com/it/personalize/faqs/
  - AMAZON LEX: https://aws.amazon.com/it/lex/faqs/
  - AMAZON KENDRA: https://aws.amazon.com/it/kendra/faqs/
  - AMAZON FORECAST: https://aws.amazon.com/it/forecast/faqs/
  - AMAZON COMPREHEND: https://aws.amazon.com/it/comprehend/faqs/

- ML Supervised: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861
  - Linear Regression
  - Forecasting
  - Classification

- ML Unsupervised: https://www.guru99.com/unsupervised-machine-learning.html
  - Clustering
  - Anomaly Detection
  - Topic modeling
  - Machine translation
  - Reinforcement Learning

- Sagemaker Ground Truth: https://aws.amazon.com/it/sagemaker/groundtruth/faqs/

- Datalake concepts and S3: https://aws.amazon.com/it/big-data/datalakes-and-analytics/what-is-a-data-lake/?nc=sn&loc=2
- Data cleaning and formatting:
    - Nominal and categorical data
    - Unclean data
    - Outliers
    - Missing data
    - Graphics with Pandas
    - Correlation Matrix
    - Overfitting and Underfitting
    - Data splitting
    - Simple hold out
    - K Fold
    - K Fold random shuffle
    - Stratified K Fold
    - Leave one out
- Tools for Supervised/Unsupervised analysis:
    - ML Supervised: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
        - RMSE, MSE, R2
        - Accuracy, F1 Score, Precision, Recall, TNR, AUC-ROC
    - ML Unsupervised: https://www.guavus.com/technical-blog/unsupervised-machine-learning-validation-techniques/
        - Internal validation
        - External validation
        - Twin-Sample Validation
    - ML Deep Learning: know at least some bits
- Sagemaker: Training Job, how to manage and create configurations: https://docs.aws.amazon.com/sagemaker/latest/dg/train-model.html
- Sagemaker: possible algorithm solutions
- Sagemaker: algorithms in details: https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html
    - Blazing Text

    - Deep AR
    - Factorization Machine
    - Image Classification
    - IP Insight
    - K-Means
    - KNN
    - LDA
    - Linear Learner
    - Neural Topic Model
    - Object2Vect
    - Object Detection
    - Principal Component analysis
    - Random Cut Forest
    - Semantic Segmentation
    - Seq2Seq
    - XGBoost

- Sagemaker: tuning: https://towardsdatascience.com/demystifying-model-training-tuning-f4e6b46e7307

    - Feature extraction
    - Numeric Transformation
    - Binning
    - Scaling
    - Categorical or nominal data
    - How to manage parameters
    - How to manage hyperparameter

- Sagemaker: deploy: https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html

    - Deploy by yourself
    - Deploy with Sagemaker
    - How to deploy
    - Blue/Green Deploy: https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/bluegreen-deployments.html
    - A/B Deploy:

https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/ab-testing.html

  - Canary Deploy: https://docs.aws.amazon.com/whitepapers/latest/wellarchitected-machine-learning-lens/canary-deployment.html

- Batch Inference: https://docs.aws.amazon.com/sagemaker/latest/dg/batch-transform.html
- Online Inference: https://docs.aws.amazon.com/sagemaker/latest/dg/inference-pipeline-real-time.html
- Online Vs Batch: https://mlinproduction.com/batch-inference-vs-online-inference/
- Sagemaker: logging and Concept Drift: https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-overview.html
- Kinesis Data Stream VS Firehose: https://aws.amazon.com/it/kinesis/data-streams/faqs/?nc=sn&loc=5

https://aws.amazon.com/it/kinesis/data-firehose/faqs/

Cheats, tips & tricks

1. Typically, in the case of **Simple Hold Out,** standard values are **80/10/10** or **70/15/15**.
2. If in some questions you find something related to **historical data** you'll have to work with **supervised algorithms**.
3. What are the benefits of **Random** and **Bayesian Optimizer** for hyperparameters in Sagemaker compared to **Grid Optimizer or manual? They better explore parameters space better verifying unusual combinations.**
4. Random optimizer is faster than Bayesian, but the second one is more precise.
5. Confusion matrix can also be **NxN in dimension**.
6. Term Frequency – Inverse Document Frequency: a high value means a rare term.
7. Oversampling and undersampling **are not useful for regression (both logistic and linear)**
8. When we have a strong unbalanced class, **class probability threshold can be put from 0.5 to a higher value for the unbalanced class.**
9. Oversampling **must be done after splitting to avoid data bleeding.**
10. To **exit from local minima** you need to reduce **batch size** and **also learning rate to reduce oscillatory effect of a small batch size promoting a better convergence.**
11. For oversampling a basic approach in questions is: **GAN is better than SMOTE which is better than manual.**
12. When we have missing values **a viable solution is applying machine learning techniques to replace values**, in general, is a good approach.
13. When I have questions regarding **recommendation engine or in general how to predict users' choices** if we have **collaborative filtering** as a possible answer than it's usually a good answer.
14. If you must decide: **AWS Glue is more batch oriented, Kinesis Firehose is for streaming real time data.**

15. If you have forecasting climate problems, you generally have to use **Linear Regression (if there aren't more fitting algorithms, this is usually sufficient)**.
16. Firehose input lambda has a timeout of **3 sec, so if you want to execute longer and complex input operations, this value is not sufficient and must be increased.**
17. Scatter Plot is best used for 2 dimension analysis, Histograms for 1 dimension.
18. Logistic Regression: **is used for binary classification, don't be fooled by its name!**
19. NTM: neural topic model: is used to group topics into categories.
20. Semantic Segmentation: helps to find high level features by making a per pixel analysis of an image.
21. Firehose is not particularly useful if data is already in S3, it takes data from a stream!
22. For deep learning and overfitting problems, use these schemas as a valid reference:

    < DROPOUT + > REGULARIZATION + > FEATURE INCREASE = < OVERFITTING

    Sono i parametri da modificare per ridurre l'overfitting. Inoltre:

    > DROPOUT = > NOISINESS

23. When you're using Kinesis Data Stream there is a special variant **video stream** that is used **in combination with Rekognition Video in a simple and clear way**. There is a specific use case in AWS guides.
24. AWS Comprehend **is not only** for **Sentiment Analysis** but also for **Topic Analysis**, **Key Extraction,** and **Entity Recognition**, also it uses NLP, so in case you'll have the same topics in a question, you can better understand if you need to use NLP.
25. L1 and L2 are regularization terms used to reduce overfitting making the algorithms more conservative when you have a dataset with many features.
26. Cross Validation is based on hyperparameters, it helps with overfitting when dataset is small, so when risk is higher due to low quantity of data.
27. Deep Learning: substitute a model with a neural network for analysis.
28. Deep Learning features can be analysed as raw because in complex network, layers themselves are able to transform low level features in higher ones.
29. In Deep learning we have different types of **activation function**:

    1. Sigmoid
    2. ReLU
    3. Softplus
    4. TanH

30. When we talk about Deep Learning **CNN** or Convolutional Neural Network (very good for image processing) and **Long Short Term Memory,** are the most used for versatility and scalability.

Questions: where to find them?

In general, because of AWS' NDA about exam's questions, its difficult to obtain **official** material (or at least officious and reliable)...This list of links points to resources personally selected by me as proven to be verified, trustable and functional to certification's preparation:

- **Official AWS questions (free):** https://d1.awsstatic.com/training-and-certification/docs-ml/AWS-Certified-Machine-Learning-Specialty_Sample-Questions.pdf
- **Official AWS test exam (paid):** https://www.aws.training/certification?rightcta=mlexam
- **Exam readiness class (paid):** https://www.aws.training/SessionSearch?pageNumber=1&courseId=38153
- **Certbolt test exam with 65 questions** – only on Windows system or VM – **highly recommended (free)**: https://www.certbolt.com/aws-certified-machine-learning-specialty-exam-dumps
- **Some studying materials from Udemy (paid)**: https://www.udemy.com/topic/aws-certified-machine-learning-specialty/
- Various questions: found on Google by searching "**aws machine learning specialty dumps**". They can be useful to verify your knowledge but **not recommended as many of them are unfortunately not correct**. Use them just as a proof of your own knowledge.

Side note: remote exam with Pearson/VUE

The remote exam involves accessing the Pearson/VUE panel and registering for a date when online proctoring is available, necessary to ascertain the validity of the exam. Given the emergency situation this year, it is advisable to book the session **well in advance**, to be sure of finding a convenient date and time.

You must have a valid ID card or **driving license or passport**, of which clear and understandable photos will be required. A damaged document or an illegible photo will cause significant delays in the scheduling due to telephone inquiries.

Software available for all Windows, Mac and Linux platforms will be downloaded, which will be used to take the exam and which will verify the room in which the test will take place by means of photos from the mobile phone. Also in this case it is necessary to scrupulously follow the instructions received via email upon registration, to carry out the validation tests before the exam date.

During the examination, a proctor will follow you from your webcam and the software downloaded will run in fullscreen to prevent the use of other programs.

Should any problem occur, the proctor will notify you via internal chat.

To conclude, if you were wondering if it was possible to prepare the AWS Certified Machine Learning Specialty – or an AWS specialty in general – on your own, now you know that the answer

is ... **technically yes** 🙂

However, for those who are beginners with AWS certifications or those who need structured training, it is important to know that **AWS puts official courses at your disposal** by delivering them through recognized APN Training Partners. **Amazon Web Services has selected beSharp as one of the few APN Training Partners authorized to provide official AWS courses.** Our training classes are developed and managed by AWS multi-certified Cloud Experts to ensure that content reflects the latest best practices. The classroom training offers participants the opportunity to receive live feedback and answers to questions directly from an expert instructor. They also include practical **workshops** designed to consolidate your knowledge.

If you want to prepare this or any other official AWS certification, read all the details and contact us to plan your training.

Good luck with your next AWS certification and see you soon!

Alessandro Gaggia

Head of Software Development

Get in touch

beSharp.it
proud2becloud@besharp.it